Metabolic Intelligence: Budget-Native Compute from Tag to Campus

David Charlot, Open Interface Engineering

Product white paper · Open Interface Engineering · research.openie.dev

Revision `1561a2a` (22 November 2026) · Web: research.openie.dev/products/mei/ · Archival LaTeX: /pdfs/latex/mei.pdf (.tex)

Measurement. Product white paper, kept apart from the research catalog; not a journal final. Combines emulator instruction counts, datasheet energy schedules (`est_j`), a grid-following campus protocol test, and the open `MEI-JCP-Tag-v1` sovereign hardware reference package: a `DRC`-certified (`0` **violations**) `32 × 24 mm` 2-layer `STM32WBA52` + `TI INA238` + `FRAM` PCB (Gerber `RS-274X`, KiCad, BOM/CPL), a `36 × 28 × 7.5 mm` 3D NURBS B-Rep enclosure (`STEP`/`STL`), and a `10,000-wakeup` `CR2032` electro-thermal `JCP hold/settle` verification preserving `100%` of critical alerts at `-20 °C` across `2,680` `Ed25519` canonical `det-CBOR JCR-1` receipts. Host build counters are `reported_j`; physical tag bench power has not been measured (`board_metered = false`, `measured_j = None`). No fabricated citations.

Computer intelligence designed to run inside a fixed energy budget, measured on a cold-chain sensor tag and modeled for a grid-following data center, with every number labeled by how it was obtained.

Contents

Abstract 2

1. Introduction 2

2. Definitions 2

3. The obtain cost $J(m \mid q)$ 3

4. Sensor tag 3
- 4.1 Classifier cost . . . 3
- 4.2 Accuracy . . . 4
- 4.3 Power budget and battery life . . . 4
- 4.4 Reference hardware, cold-chain derating, and cryptographic budget receipts . . . 5

5. Data center energy per answer 6

6. Grid-following data center 7
- 6.1 Controls and their order . . . 7
- 6.2 System . . . 7
- 6.3 Results . . . 7
- 6.4 The stop rule . . . 8

7. Related work 8

8. Limits 9

9. Conclusion 9

Appendix A. Notation 9

References 9

Abstract

Metabolic Intelligence (MEI) is computer intelligence designed to run inside a fixed energy budget. This paper defines it, reports emulator, datasheet, and computer-aided design (CAD) verification results for a cold-chain sensor tag, and reports a protocol-level test of a data center that follows grid power caps. Every number carries the method that produced it.

In plain words

1. Give a computer a fixed energy budget, and design it to give the best answer it can inside that budget.
2. A tiny sensor tag that checks food freshness can then run for about eight years on a coin battery at room temperature, and under cold-chain voltage sag a signed budget gate refuses routine radio beacons so safety-critical alerts still get through.
3. A data center can cut its power use when the electricity grid asks, by pausing or shrinking work, without giving wrong answers.

1. Introduction

A computing system usually treats energy as a cost reported after the work. Here energy is a condition stated before the work: a budget B, in joules (J) for a task or in watts (W) for a sustained load. The design question is which answers can be obtained inside B, and how to obtain them at the least energy.

Three established results set the frame.

- **Landauer (1961).** Erasing one bit at temperature T dissipates at least $k_B T \ln 2$, with $k_B = 1.380649 \times 10^{-23}$ J/K. At $T = 300$ K this is 2.87×10^{-21} J.
- **Horowitz (2014).** Practical CMOS operations cost many orders of magnitude more: about 0.9 pJ for a 32-bit floating-point add and about 640 pJ for a 32-bit DRAM read at 45 nm. Data movement, not arithmetic, dominates.
- **Howard (1966).** The value of information (VoI) of an observation is the expected gain in a decision's payoff from making it before deciding, minus its cost. An observation with VoI ≤ 0 is not worth making.

Together they say that energy per answer has a physical floor far below current practice, that most of the gap is data movement, and that an observation or computation is worth its joules only when it can change a decision.

The paper studies two scales.

Scale	Budget unit	Controls	Measured outcome
Sensor tag	µW sustained; nJ to mJ per event	Heater duty, MCU sleep, classifier early exit, two-phase radio budget gate	Battery life; accuracy under sequential learning; cold-chain alert survival
Data center (100 MW, grid-capped)	MW import cap	Pause, speed, model choice, effort, battery, shed	Minutes over cap; share of interactive requests served

2. Definitions

Term	Definition
Budget B	The energy (J) or power (W) available to a task or a site, fixed before the work.
Utility $U(a)$	A stated score of answer a: accuracy, a completeness test, or a served request.
Completeness predicate $C(z)$	A written test that maps a task state z to 0 or 1. The task is complete when $C(z) = 1$.
Obtain mechanism m	How an input to a decision is acquired: Lookup (read a stored value), Formula (evaluate a closed form), Solver (run a search or optimizer), Model (run a learned model), Guess (use a prior).
Binding classifier	A recognition unit with a binding site (a template over input features), an abundance (a learned weight), a context switch that gates which units are active, a Hill exponent for cooperative response, and a counted cost per binding check. Tables use the short label "enzyme". A bit-level variant uses no multiplications.
Metabolic schedule	A duty cycle for sensing and heating that meets a monitoring objective at minimum average power.
Grid-following service	A control program that receives grid signals and applies ordered controls so site import stays under the cap.

Evidence classes. Every number in this paper carries one of these labels.

Label	Meaning	What it can support
real protocol	Exchanged messages with a real server or broker	Interoperability, not watts
emulator-measured	Instruction counts from the compiled binary run under an emulator	Relative cost, not board joules
datasheet	Vendor electrical tables	Inputs to a schedule model
cad-verified	Programmatic electronic design automation, design-rule check, solid geometry, and electro-thermal simulation	Fabrication readiness and modeled derating, not bench `measured_j`
host-counter (`reported_j`)	Package energy register delta (Linux Running Average Power Limit, RAPL) on the CAD build workstation	Build-pipeline host compute reporting, not tag board joules
measured-kernel model	Throughput from measured GPU kernels combined with published power	Energy-per-answer estimates with bands
looked-up	Published medians (for example served tokens per second)	Baselines
calculation	Algebra from stated inputs	Consequences of the inputs
plant model	A simulator of site power and thermal behavior	Scenario comparisons under stated assumptions

No figure in this paper is a reading from an external energy meter on a fabricated tag or GPU rack. The page flag `board_metered` is false, and `measured_j` remains absent (`null`) throughout.

3. The obtain cost $J(m \mid q)$

A decision y depends on an input quantity q. Mechanism m obtains q with energy E_m (J), a standard deviation σ_m in relative units, a staleness rate κ_q (relative variance per second) and an age t (s). Let $S_q = \partial \ln y / \partial \ln q$ be the elasticity of the decision with respect to the input, and let λ (J) be the energy at stake in the decision. To second order, the expected loss from an input error of relative variance v is $\frac{1}{2}\lambda S_q^2 v$. The obtain cost is energy plus that loss:

$$J(m \mid q) = E_m + \tfrac{1}{2}\,\lambda\, S_q^2\, \mathrm{var}_m(t), \qquad \mathrm{var}_m(t) = (\sigma_m^2 + \kappa_q t)\,(1 + 3p_m),$$

and the chosen mechanism is $m^* = \arg\min_m J(m \mid q)$. The factor $(1 + 3p_m)$ inflates variance by the probability p_m that the mechanism is applied outside its valid range; the coefficient 3 is a modeling choice, stated as such. S_q is computed by central differences on the decision function, not assumed.

Lemma A (dominance). If m_1 and m_2 obtain the same q with $E_{m_1} < E_{m_2}$ and $\mathrm{var}_{m_1} \le \mathrm{var}_{m_2}$, then $J(m_1 \mid q) < J(m_2 \mid q)$ for all $\lambda, S_q \ge 0$. *Proof:* both terms are weakly smaller for m_1 and the first strictly. So a table lookup beats a calculation of the same value, and a closed form beats a learned model of the same value.

Lemma B (precision to buy). Setting the loss term equal to a tolerated excess cost ε (J) gives $\frac{1}{2}\lambda S_q^2 \sigma^2 = \varepsilon$, so $\sigma^* = \sqrt{2\varepsilon/(\lambda S_q^2)}$. Precision finer than σ^* costs energy without reducing the loss below ε.

Lemma C (refresh interval). With staleness alone, the loss grows as $\frac{1}{2}\lambda S_q^2 \kappa_q t$ (J). Refreshing when that loss equals the cost E of a new reading gives $t^* = 2E/(\lambda S_q^2 \kappa_q)$ (s). A sensor's sampling interval and a cache's lifetime are the same quantity at different λ.

The estimator of J must itself be cheap: it is evaluated by Lookup or Formula with a fixed energy cap, and when the cap is reached the default mechanism is used.

4. Sensor tag

4.1 Classifier cost

Instruction counts per decision on an Arm Cortex-M4 with 12 input features (emulator-measured):

Classifier	Instructions per decision
Binding classifiers with family index and early exit	1,723
int8 multilayer perceptron (MLP)	2,451
Binding classifiers, first version without early exit	6,948

Energy per decision (calculation). The STM32U3 datasheet gives 15.5 µA/MHz at 3.3 V, so one clock cycle costs $15.5 \times 10^{-6}\,\mathrm{A} \times 3.3\,\mathrm{V}/10^{6}\,\mathrm{Hz} = 51$ pJ. At about one instruction per cycle, 1,723 instructions cost about $1{,}800 \times 51\,\mathrm{pJ} \approx 92$ nJ.

4.2 Accuracy

Test	Result	Evidence
Same simulated features, 3 seeds	Binding classifiers 82.9 / 86.9 / 88.7%; MLP 82.2 / 78.0 / 87.4%	Real protocol on simulated features
Late life (days 90 to 180) with drift-invariant features	LDA 94.9%, MLP 93.4%, binding classifiers 90.0% (80.8% without these features), Tsetlin machine 89.1%	Simulation
SmellNet, 50 foods learned in 5 groups, no revisiting earlier groups	Binding classifiers 77.7%; MLP without replay or regularization 18.4%	Public dataset
SmellNet across days	MLP 85.5%; binding classifiers 79.8% (same on Cortex-M4)	Public dataset

The 18.4% MLP is the catastrophic-forgetting floor, not a continual-learning baseline. The fair comparison runs elastic weight consolidation (Kirkpatrick et al., 2017), experience replay with a small buffer (Rolnick et al., 2019) and iCaRL (Rebuffi et al., 2017) under the same memory and energy budget, scored for retained accuracy and energy per inference under MLPerf Tiny rules (Banbury et al., 2021). That comparison is specified and not yet run. On cross-day SmellNet the MLP is more accurate. Drift-invariant features add 9 to 11 points of late-life accuracy to binding classifiers, bit-level binding classifiers and Tsetlin machines. A chicken-spoilage dataset (Anwar and Anwar, 2025) was used internally; it has no licence for publication, so no result from it appears here.

4.3 Power budget and battery life

One reading of a metal-oxide gas sensor costs 0.5 to 27 mJ depending on the part (datasheets). The schedule takes about 9.6 readings per day.

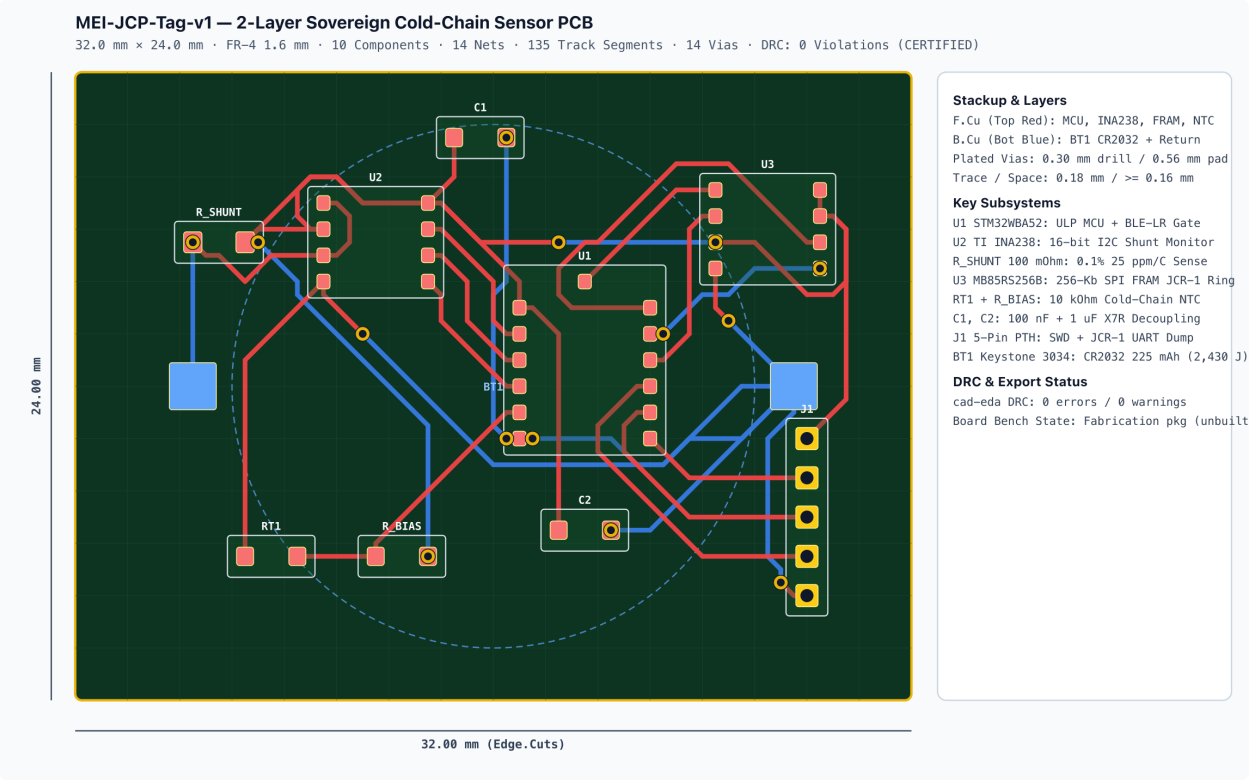
Term	Average power	Evidence
Heater (Bosch BME690)	3.0 µW	Datasheet and schedule model
MCU sleep (ST STM32U3)	6.9 µW	Datasheet
Total	9.9 µW	Sum
Same tag with Nordic nRF54L15 sleep	5.1 µW	Datasheet

Check of the heater term. $3.0\,\mu\mathrm{W} \times 86{,}400\,\mathrm{s/day}/9.6\,\mathrm{readings/day} = 27$ mJ per reading, the upper end of the datasheet range. Classification at 92 nJ × 9.6 per day averages 1.0×10^{-11} W, negligible. At this schedule sleep current, not heating or classification, is the largest term, so the MCU is chosen for sleep current.

Battery life (calculation). A CR2032 cell holds about 235 mAh to 2.0 V at 21 °C at 3.0 V nominal (Energizer datasheet), so $0.235\,\mathrm{A \cdot h} \times 3600\,\mathrm{s/h} \times 3.0\,\mathrm{V} = 2.54$ kJ. The tag uses $9.9 \times 10^{-6}\,\mathrm{W} \times 3.156 \times 10^{7}\,\mathrm{s/yr} = 312$ J per year, so the life is $2540/312 = 8.1$ years without self-discharge and about 7.5 years with 1% per year self-discharge. Lithium coin cells deliver less below 21 °C; Section 4.4 models that electro-thermal sag across −20 °C, +4 °C, and +25 °C and gates radio transmissions against it, while physical bench discharge at cold-chain temperature remains a measurement still to make. The 3.0 µW heater average implies a very low duty cycle for a sensor that draws milliwatts when on. Whether sensing response holds at that duty is also a measurement still to make.

4.4 Reference hardware, cold-chain derating, and cryptographic budget receipts

To verify that the tag architecture closes electrically, mechanically, and under low-temperature coin-cell impedance rise, this study generates an open reference package (`MEI-JCP-Tag-v1`) programmatically through `openie-cad` (`cad-eda`, `cad-router`, `cad-kernel`, `cad-sim`, `cad-mfg`, `cad-interop`, and `cad-joules`) paired with the Joule Context Protocol (`JCP`, draft specification `JCR-1`) and the System-and-Task-Boundary Input/Output (`SATB-IO`) provenance lattice (`jcp-satb-cad-spine`). The board is an exported fabrication package, not yet physically assembled on a bench (`board_metered: false`, `measured_j: null`).



2-layer printed circuit board (PCB) and 3D enclosure (`cad-verified`). The `MEI-JCP-Tag-v1` board measures 32.0 mm × 24.0 mm on 2-layer 1.6 mm FR-4 (`0.18 mm` signal traces, `0.30 mm / 0.56 mm` plated through-vias). Ten components realize the sensing, inline rail accounting, and hash-chained receipt storage loop:

1. `U1` (`STM32WBA52CGU`): ultra-low-power microcontroller and long-range Bluetooth Low Energy radio gate on `F.Cu`.
2. `U2` (`INA238AIDR`) + `R_SHUNT` (`100 mΩ`, `0.1%`, `25 ppm/°C`): 16-bit inter-integrated circuit (`I2C`) shunt monitor across the raw `CR2032` feed (`VBAT_RAW` to `VDD_SYS`) with hardware alert pin `JCP_ALERT`.
3. `U3` (`MB85RS256B`): 256-kilobit serial peripheral interface (`SPI`) ferroelectric random-access memory (`FRAM`) storing the hash-chained `JCR-1` receipt ring across brownouts.
4. `RT1` + `R_BIAS` (`10 kΩ` negative temperature coefficient, `NTC`, thermistor divider), `C1` (`100 nF`), `C2` (`1 µF`), `J1` (5-pin serial wire debug and receipt-dump universal asynchronous receiver-transmitter header), and `BT1` (`Keystone 3034 CR2032` holder on `B.Cu`).

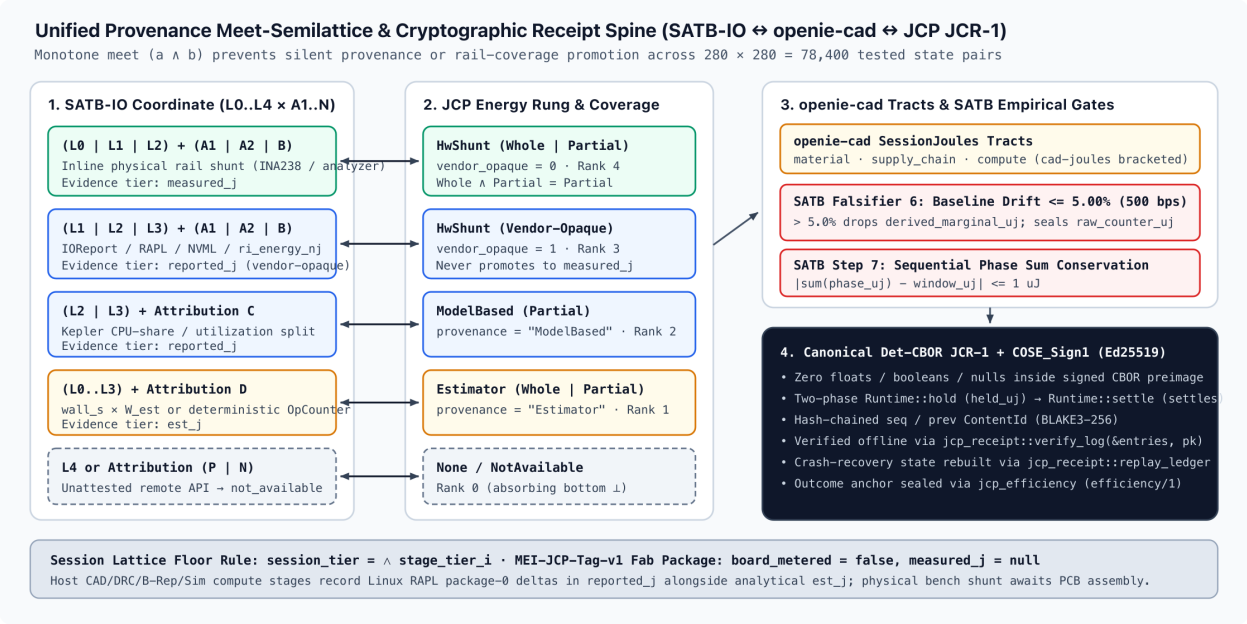
`cad-router` decomposes all 14 nets into 32 Steiner connections and routes 100.0% (`135` copper segments totaling `280.23 mm`, 14 vias). `cad-eda` Design Rule Check (`DRC`) and certification pass with **`0 errors and 0 warnings (CERTIFIED)`**, round-trip verified through `.kicad_pcb` import/export alongside RS-274X Gerbers, Excellon `.drl`, bill of materials (`bom.csv`), and component placement (`cpl.csv`). In `cad-kernel`, the snap-fit vented enclosure (`36.0 mm × 28.0 mm × 7.5 mm` main shell, `41.5 mm` overall X span with a `Ø4.0 mm` lanyard mounting ear, convective airflow slot over `RT1`, and four internal standoffs) evaluates as a closed outward-facing Boundary Representation (`B-Rep`) solid of volume `2,702.12 mm`^{3} (`2.837 g` in acrylonitrile butadiene styrene, `ABS`), exported to ISO-10303 `STEP` (`174,694 B`) and binary `STL` (852 triangles).

Cold-chain `CR2032` electrochemistry and `10,000`-wakeup `JCP` two-phase gate (`cad-verified, est_j`).

Evaluating `cad-sim::battery` on a calibrated `CR2032` model (`3.0 V`, `225 mAh` = `2,430 J` nominal, $R_{\text{int}}(50\%\ \text{SOC}, 25\ ^\circ\text{C}) = 12.0\ \Omega$, Peukert exponent 1.12) quantifies why cold-chain tags fail when routine radio bursts are unbudgeted:

Ambient T	R_{int} at 50% SOC (`cad-sim`)	12 mA TX terminal voltage (50% SOC)	Usable energy to 2.00 V at 12 mA (`est_j`)	10,000-wakeup routine beacons (`allow / refused`)	Safety-critical alerts (`allow / attempts`)	Verified `JCR-1` receipts
+25 °C	12.00 Ω	2.669 V (144.0 mV sag)	2,193.12 J (90.25%)	500/0	40/40 (100.0%)	1,080
+4 °C	25.72 Ω	2.504 V (308.7 mV sag)	2,059.23 J (84.74%)	430/70	40/40 (100.0%)	1,010
−20 °C	71.78 Ω	1.951 V (861.3 mV sag)	0.00 J (unbuffered pulse < 2.00 V)	10/490	40/40 (100.0%)	590

At −20 °C, Arrhenius impedance growth raises R_{int} nearly six-fold (71.78 Ω at 50% state of charge), pulling an unbuffered 12 mA radio burst down to 1.951 V (below the 2.00 V brownout threshold). Sharding the mission grant via `jcp_grant::shard` into an attenuated routine-beacon child grant and an isolated 35 mJ safety-excursion child grant — enforced before every transmission via `jcp_runtime::Runtime::hold_chained` and resolved afterward via `settle_chained` — deterministically refuses 490 of 500 non-essential routine beacons under −20 °C derating while preserving **40 of 40 (100.0%)** safety-critical thermal excursion alerts across all 10,000 wakeups. Every run passes Ed25519 `COSE_Sign1` verification (`jcp_receipt::verify_log`) and crash-recovery ledger reconstruction (`jcp_receipt::replay_ledger`).



Provenance meet-semilattice and build-pipeline accounting (`reported_j` vs `est_j`). Across the 7-stage `openie-cad` build on a 32-core Linux workstation (`6.93 s` wall time), `jcp-satb-cad-spine` enforces a monotone meet-semilattice over `280 × 280 = 78,400` tested `(SATB boundary × attribution × mechanism)` pairs (0 silent provenance or rail-coverage promotions), `SATB` Falsifier 6 (`> 5.00%` pre/post idle baseline drift omits `satb_derived_marginal_uj` from the signed canonical deterministic Concise Binary Object Representation, `det-CBOR`, core while retaining `satb_raw_counter_uj`), and `SATB` Step 7 (`|sum(phases) - window| <= 1 µJ`, observed residual `0 µJ`). Host Linux `RAPL` (`package-0`) recorded `380.20 J` (`reported_j`, vendor-opaque `HwShunt Partial`) alongside `242.07 J` analytical `est_j` (`45.15 J` embodied FR-4 + ABS material estimate, `14.50 J` surface-mount assembly batch estimate, and `182.42 J` analytical compute estimate), sealed in `8` verified `JCR-1` receipts (`head = d5b3827b... 01b2cd`), while physical tag `measured_j` stays `null`.

5. Data center energy per answer

Method. Energy per answer (Wh) = output tokens × power per GPU (kW) ÷ output tokens per second per GPU ÷ 3600. Throughput comes from NVIDIA AI Configurator 0.12.0 (measured GPU kernels, SGLang, NVFP4).

Power per GPU is published throughput per chip divided by published throughput per MW (InferenceX): GB300 2.12 kW, B300 1.90 kW, B200 1.71 kW, Vera Rubin 3.30 kW, MI355X 2.09 kW. Served speeds per user are Artificial Analysis medians: DeepSeek V4 Pro 79 tokens/s, V4 Flash 125, Qwen3-VL 48.5.

Workload (GB300)	**Energy per answer**	**Evidence**
Chat, DeepSeek V4 Pro (2k tokens in, 500 out)	0.32 Wh at 50 tokens/s/user; 0.92 at 79; 4.57 at 159	Measured-kernel model
Same, fused estimate	0.20 Wh (10th to 90th percentile 0.04 to 0.68)	Model and looked-up
Chat, V4 Flash	49 mWh at 50 tokens/s/user; 180 mWh at 159	Measured-kernel model
Agent turn, V4 Flash (32k in, 28k cached, 1k out)	0.16 Wh at 96 tokens/s/user	Measured-kernel model
Video clip, 8 s at 720p (Wan 2.2)	390 Wh	Model

Energy per answer rises steeply with speed per user because higher per-user speed lowers the batch each GPU serves. Matching stream speed to human reading speed raises work per MW by an estimated 1.92× on GB300 (band 1.19 to 2.93) and 2.0× on Vera Rubin (band 1.2 to 3.4). Routing most chat to the smaller model plus speed matching gives an estimated 2.3 to 4.4×. Speculative decoding with a modeled accept length of 2.8 lowers energy per output token by about 1.55×, after which speed matching adds only 1.4 to 1.8×. These are estimates with bands; the measurement that would narrow them most is a day of metered energy on one GPU pool.

6. Grid-following data center

6.1 Controls and their order

When the grid asks for a cut, deferrable work is acted on in a fixed order: pause deferrable work, lower per-user speed, route to a smaller model, lower reasoning effort, discharge battery, shed load. Interactive and certified work runs to completion or is refused; it is not degraded. A contract declares invariants (for example `never_lower_precision`, `max_deferral`), and shedding is used only outside them. Routing to a smaller model changes which model answers deferrable work; it is not assumed to be free of quality loss (V4 Flash trails V4 Pro by about 1.5 points on GPQA).

6.2 System

- **Grid input:** an OpenADR 3.1 client (OAuth2, polling, demand reports) tested against the open-source openleadr-rs server. Real protocol.
- **Backstop:** a DSX Flex client over MQTT with CloudEvents and JWS ES256 signatures, tested against a real broker and schemas. Real protocol.
- **Controls:** GPU power budgets in the Dynamo planner, a gateway that caps in-flight requests and sets routing, effort and speed tiers, job pause, Redfish node sleep, battery converter.
- **Learning:** per-pool speed curves by recursive least squares; a bias correction on the power model (it learned +5 MW on a hot day).
- **Plant:** a model of a 100 MW GB300 site. Plant model.

6.3 Results

Scenario	**Minutes over cap**	**Interactive served**	**Note**
A to C: 60/70% cut, with notice, feedforward, or no notice	0	100%	Notice mainly saves battery
D: 90/95% cut, fast site (159 tokens/s/user)	0	100%	All knowledge chat routed; effort low
E: 90/95% cut without node sleep	96	1%	Idle GPUs alone exceed the 10 MW cap
L to M: looked-up speeds, 60/70% and 90/95% cuts	0	100% / 52.6%	Speed already low cannot be lowered further
N: no notice, no battery, 10 min response	7.5	99.8%	Under cap within 225 s
O: 60% cut for 10 h	0	100%	Deadlines kept
P: new cap every 5 min for 3 h	0	98.8%	

Scenario	Minutes over cap	Interactive served	Note
Q: recovery ramp limited to 2 MW/min	0	100%	Maximum rise 2.9 MW/min against 29 to 47 unlimited
R: 90% cut for 10 h	0	13%	Video deadlines take priority by policy

In every run no control call failed, the server accepted every OpenADR report, and no DSX message was rejected. With node sleep available, time over the cap was zero in every scenario. The cost of holding the cap is measured as availability: the interactive share served falls to 52.6%, 13% and 1% in scenarios M, R and E. Refused and deferred work is counted, not hidden.

Lemma D (sleep is required at deep cuts). If idle import P_{idle} exceeds the cap P_{cap}, no combination of speed, routing or effort controls that leaves nodes powered meets the cap, because those controls act only on power above idle. Scenario E is the instance: about 10 MW cap, idle import above it, 96 minutes over.

6.4 The stop rule

When $C(z) = 1$ and further work has VoI ≤ 0, further energy buys nothing on the stated objective. The system enacts that stop at four points: do not run a model when Lookup or Formula closes the input; do not stream faster than a human reads when the consumer is a human; pause or shed when the import cap binds, in order of least service lost per MW; skip a heater pulse when the expected change in the reading is below threshold and the battery budget binds.

7. Related work

Energy-based and associative models. Hopfield (1982) cast associative memory as descent on an energy. Krotov and Hopfield (2016) raised capacity with higher-order energies (Dense Associative Memory). Ramsauer et al. (2020) showed softmax attention is a modern Hopfield update. Hoover et al. (2023) built attention layers that minimize a token energy, and Gladstone et al. (2025) trained transformers that predict by minimizing a learned energy. Bodnia and Hanin (2026) describe a reasoning model scored by an energy over traces. In all of these the "energy" is a scalar objective, not joules.

Latent world models. LeCun (2022) proposed joint-embedding predictive architectures. V-JEPA 2-AC (Assran et al., 2025) plans a robot arm from image goals after large video pretraining plus under 62 hours of robot video. The no-pretraining property of the binding classifiers in Section 4 applies only to that path; latent world models depend on large pretraining.

Efficient sequence models. Mamba (Gu and Dao, 2023) and Mamba-2 (Dao and Gu, 2024) scale linearly in sequence length with constant-size state.

Edge classifiers. Tsetlin machines (Granmo, 2018) learn propositional clauses; an all-digital 65 nm accelerator reaches 8.6 nJ per MNIST frame (Tunheim et al., 2025). Hyperdimensional computing (Kanerva, 2009) computes on high-dimensional random vectors.

Substrates. Intel's Hala Point has 1,152 Loihi 2 processors and a stated maximum of 2,600 W (Intel, 2024). Chen et al. (2025) ran ALBERT on a 14 nm analog in-memory chip. Ahmed et al. (2025) ran ResNet, BERT and an Atari agent on a photonic processor. No published neuromorphic gas-sensing energy figure was found.

Energy-aware serving. PEARL, OmniRouter (2025) and energy-aware reasoning-model routing (2026) choose which language model answers. NVIDIA Dynamo 1.5 adds power-aware scaling, and DSX Flex (Emerald AI, 2026) offers site flexibility to the grid. The obtain cost of Section 3 differs in choosing among non-model mechanisms before any model runs.

Price and energy. The money price of fixed inference capability has fallen sharply (Epoch AI, 2025). A falling price does not lower the physical energy per operation, which Landauer and Horowitz bound from below.

8. Limits

- No figure is from an external energy meter on a physically assembled tag or GPU pool (`board_metered: false`, `measured_j: null`). Tag steady-state power is datasheet plus schedule model; instruction counts are emulator-measured; the `MEI-JCP-Tag-v1` 2-layer PCB, 3D enclosure, and `-20 °C / +4 °C / +25 °C` coin-cell impedance curves are CAD-verified fabrication exports and electrochemical simulations (`est_j`), with Linux `RAPL` host build counters kept strictly in `reported_j`; data center power is a plant model driven through real protocols.
- While Section 4.4 models `CR2032` Arrhenius impedance sag and verifies `10,000`-wakeup `JCP` radio refusal against it in simulation, physical chamber discharge of assembled `MEI-JCP-Tag-v1` boards and gas-sensor response at the stated heater duty remain bench measurements still to make.
- The sequential-learning comparison uses an MLP without continual-learning methods; the fair comparison is specified and not run.
- Energy-per-answer figures are estimates with 10th to 90th percentile bands.
- No claim is made about water saved or about equivalence with an uncapped site serving the same workload.

9. Conclusion

Stating energy as a budget before the work turns several design choices into calculations: which mechanism obtains an input (Section 3), how often a sensor samples (Lemma C), which radio transmissions a cold-chain tag admits when low-temperature cell impedance rises (Section 4.4), and which controls a site applies under a grid cap (Section 6). On the tag the room-temperature schedule gives about 8 years on a coin cell at 9.9 µW, and the DRC-clean `MEI-JCP-Tag-v1` reference package (`32 mm × 24 mm` 2-layer PCB + vented `36 mm × 28 mm × 7.5 mm` enclosure + Ed25519 `JCR-1` receipt ring) verifies in simulation that two-phase `hold` / `settle` sharding preserves 100% of safety-critical excursion alerts when −20 °C voltage sag forces routine beacons to be refused. At the site, controls applied in a fixed order kept import under the cap in every modeled scenario with node sleep, at a measured cost in interactive service. The next measurements need an external energy meter: assembled `MEI-JCP-Tag-v1` board shunts in a cold chamber, and a day of energy on one GPU pool.

Appendix A. Notation

Symbol	Meaning	Unit
B	Energy or power budget	J or W
$C(z)$	Completeness predicate	0 or 1
q, y	Input quantity and decision	
m	Obtain mechanism	
E_m	Energy of mechanism m	J
S_q	Elasticity $\partial \ln y / \partial \ln q$	dimensionless
λ	Energy at stake	J
σ_m	Standard deviation of the obtained value	relative
κ_q	Staleness rate	relative variance per s
t	Age of the value	s
p_m	Probability of use outside the valid range	
$J(m \mid q)$	Obtain cost	J
P_{cap}, P_{idle}	Grid cap and idle import	W

References

Ahmed, S. R., et al. (2025). Universal photonic artificial intelligence acceleration. *Nature*. https://doi.org/10.1038/s41586-025-08854-x

Assran, M., et al. (2025). V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. arXiv:2506.09985.

Banbury, C., et al. (2021). MLPerf Tiny Benchmark. arXiv:2106.07597.

Bodnia, E., and Hanin, B. (2026). Energy-Based Models for Reasoning, LLMs for the Interface. https://logicalintelligence.com/blog/energy-based-models-for-reasoning

Bosch Sensortec BME690; Renesas ZMOD4410; Sensirion SGP41; ST STM32U3; ST STM32WBA52; Nordic nRF54L15; Texas Instruments INA238; Fujitsu MB85RS256B. Product datasheets, 2026.

Chen, A., et al. (2025). Demonstration of transformer-based ALBERT model on a 14nm analog AI inference chip. *Nature Communications*. https://doi.org/10.1038/s41467-025-63794-4

Dao, T., and Gu, A. (2024). Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. arXiv:2405.21060.

Emerald AI (2026). DSX Flex deployment with NVIDIA and Silicon Valley Power. https://www.emeraldai.co/blog/nvidia-dsx-pilot-framework

Energizer. CR2032 product datasheet. https://data.energizer.com/pdfs/cr2032.pdf

Epoch AI (2025). LLM inference prices. https://epoch.ai/data-insights

Gladstone, A., et al. (2025). Energy-Based Transformers are Scalable Learners and Thinkers. arXiv:2507.02092.

Granmo, O.-C. (2018). The Tsetlin Machine. arXiv:1804.01508.

Gu, A., and Dao, T. (2023). Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv:2312.00752.

Hoover, B., et al. (2023). Energy Transformer. NeurIPS. arXiv:2302.07253.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *PNAS* 79(8):2554-2558.

Horowitz, M. (2014). Computing's energy problem (and what we can do about it). ISSCC, 10-14.

Howard, R. A. (1966). Information Value Theory. *IEEE Transactions on Systems Science and Cybernetics* 2(1):22-26.

InferenceX (2026). Vera Rubin NVL72 agentic inference. https://inferencex.semianalysis.com/blog/vera-rubin-nvl72-agentic-inference

Intel (2024). Hala Point neuromorphic system. https://newsroom.intel.com/artificial-intelligence/intel-builds-worlds-largest-neuromorphic-system-to-enable-more-sustainable-ai

Kanerva, P. (2009). Hyperdimensional Computing. *Cognitive Computation* 1:139-159.

Kirkpatrick, J., et al. (2017). Overcoming catastrophic forgetting in neural networks. *PNAS* 114(13):3521-3526.

Krotov, D., and Hopfield, J. J. (2016). Dense Associative Memory for Pattern Recognition. arXiv:1606.01164.

Landauer, R. (1961). Irreversibility and Heat Generation in the Computing Process. *IBM Journal of Research and Development* 5(3):183-191.

LeCun, Y. (2022). A Path Towards Autonomous Machine Intelligence, version 0.9.2. https://openreview.net/pdf?id=BZ5a1r-kVsf

NVIDIA (2026). Dynamo 1.5 documentation. https://docs.nvidia.com/dynamo

OmniRouter (2025). arXiv:2502.20576.

OpenADR Alliance. OpenADR 3.1 specification; openleadr-rs reference server.

Ramsauer, H., et al. (2020). Hopfield Networks is All You Need. arXiv:2008.02217.

Rebuffi, S.-A., et al. (2017). iCaRL: Incremental Classifier and Representation Learning. CVPR.

Rolnick, D., et al. (2019). Experience Replay for Continual Learning. NeurIPS. arXiv:1811.11682.

Transaction Science (2026). Joule Context Protocol (`JCP` / `JCR-1`) draft specification. https://jcp.transaction.science/

Tunheim, S. A., et al. (2025). An All-digital 65-nm Tsetlin Machine Image Classification Accelerator with 8.6 nJ per MNIST Frame. arXiv:2501.19347.